Populating CSV Files from Unstructured Text with LLMs for KG Generation with RML

Jan Maushagen[1], Sara Sepehri[2], Audrey Sanctorum[1], Tamara Vanhaecke[2], Olga De Troyer[1] and Christophe Debruyne[3,*]

[1] *Web & Information Systems Engineering (WISE) Lab, Vrije Universiteit Brussel, Brussels, Belgium*
[2] *Research Group of In Vitro Toxicology and Dermato-Cosmetology (IVTD), Vrije Universiteit Brussel, Brussels, Belgium*
[3] *Montefiore Institute of Electrical Engineering and Computer Science, University of Liège, Liège, Belgium*

Abstract

We report on an exploratory study using Large Language Models (LLMs) to generate Comma-Separated Values (CSV) files, which are subsequently transformed into Resource Description Framework (RDF) using the RDF Mapping Language (RML). Prior studies have shown that LLMs sometimes have problems generating valid and well-formed RDF from unstructured texts, i.e., issues with RDF, not the contents. We wanted to test whether the generation of CSV led to fewer issues and whether this would be a viable option for allowing domain experts to be actively part of the Knowledge Graph (KG) population process by allowing them to use familiar tools. We have built a prototype illustrating this idea, and the results seem promising for further study. The initial prototype uses zero-shot training and is built on GPT-4. The prototype takes the unstructured text and the CSV file's structure as input and uses the latter to generate prompts to fill in the cells' values. Future work includes analyzing the effect of different prompting strategies. The limitation, however, is that such an approach only works for projects where domain experts work with spreadsheets for pre-existing mappings.

Keywords

KG Construction, LLMs, End-user Involvement

1. Introduction

Knowledge Graphs (KGs) enable us to organize, represent, and reason about structured information integrated from various sources. However, KG construction remains challenging due to the heterogeneity and complexity of real-world data sources. End-user and domain-expert involvement in all KG construction activities, such as ontology engineering, data transformation, data enrichment, and quality assurance, is a challenge requiring bespoke methods and tools, as exemplified in [1] and [2]. In [1], we proposed a method in the toxicology domain that relied on domain experts populating a set of spreadsheets, which are subsequently transformed into RDF using RML. Our approach also includes an end-user approach based on the block metaphor.

Large Language Models (LLMs) have demonstrated their potential for natural language understanding and generation tasks, and their use has been explored in KG construction. [3],

SEMANTiCS 2024: 20th International Conference on Semantic Systems, September 17–19, 2024, Amsterdam, The Netherlands

[*] Corresponding author.

0000-0003-0197-2111 (J. Maushagen); 0000-0003-1023-7987 (S. Sepehri); 0000-0002-4872-4687 (A. Sanctorum); 0000-0002-6685-7299 (T. Vanhaecke); 0000-0002-8457-7143 (O. De Troyer); 0000-0003-4734-3847 (C. Debruyne) for instance, generated RDF from unstructured text and noticed differences when the LLM was requested to produce Turtle, JSON-LD, ... LLMs are not only used to generate RDF, but their use has been explored in declarative mappings as well. In [4], the authors demonstrated that LLMs can be used to engage with RML [5] mappings and that the output (RDF, queries, etc.) is of fairly high quality. However, in [6], the authors explored various LLMs to generate RML mapping and noticed that they tended to generate syntactically correct RDF but invalid mappings.

As demonstrated in [3], applying LLMs to KG construction may still have suboptimal results. Recognizing that the generation of RDF from unstructured text has some challenges, we explored using LLMs to distill simple (i.e., CSV) semi-structured information from unstructured text that domain experts can more easily validate and refine with spreadsheets. We believe this approach would yield better results in contexts where one has an ontology and data can easily be entered into such files. This paper elaborates on our approach and reports on our initial findings.

2. Context

This study was conducted in the context of the TOXIN project. [1] A major part of this project was to gather and integrate information about *in vivo* tests, described in *Safety Evaluation Opinions*, issued by the Scientific Committee on Consumer Safety (SCCS) about cosmetic ingredients, in a knowledge graph. Each Opinion, i.e., dossier, contains information about experiments or *tests* of an ingredient (compound) on laboratory animals (the compound, quantities, exposure, animals, outcomes, ...). The data contained in these dossiers are integrated into a KG to provide more efficient access to this data for toxicologists.

Our current method for populating the knowledge graph (KG) relies on domain experts reading and interpreting safety evaluation opinions to enter the details of experiments in spreadsheets, which are subsequently transformed into RDF using RML. Our approach includes an alternative (end-user) approach based on the block metaphor to enter the data into the KG directly.

While this process ensures the authoritative nature of the data, it is inherently tedious and time-consuming. The automation of this process was hampered by the variety in structure, presentation, and even writing style (e.g., the use of negation) across opinions.

3. Approach

While LLMs have been demonstrated to be promising, the aforementioned problems regarding RDF generation are problematic if the domain experts are not knowledgeable in these technologies. The ontologies and mappings have already been engineered in the TOXIN project. We can thus explore whether a) LLMs are better at generating (CSV) tables or, at least, finding the relevant information in the text to construct such a table, and b) whether such an approach could assist domain experts in filling those spreadsheets more efficiently. To this end, we have built a prototype assistant, see Figure 2, that takes a safety evaluation opinion and the table's structure as input.

In the current prototype (Figure 1), the text about the experiments (or studies) is extracted using regular expressions (1), and the column headers are used to generate the prompts (2). The

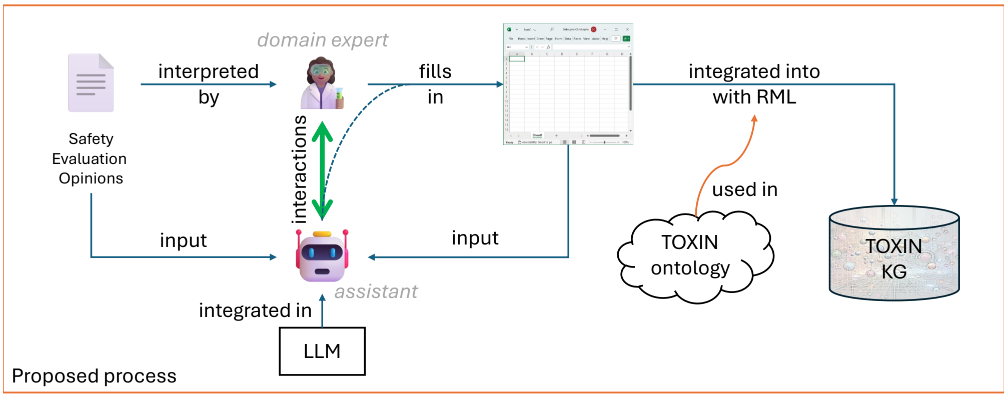
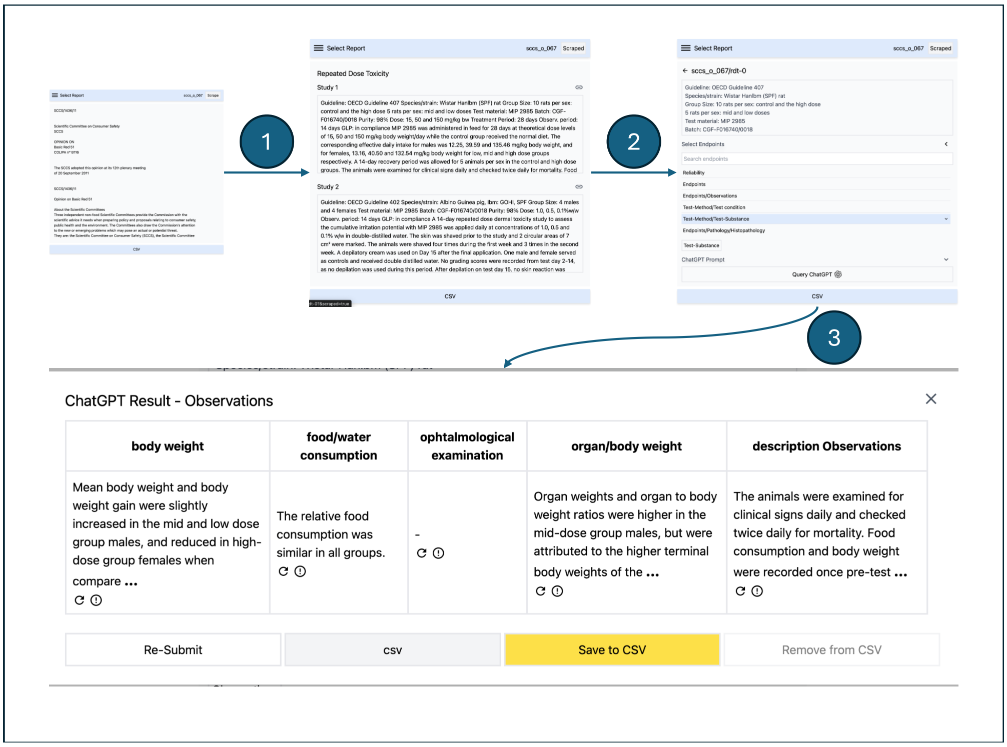


Figure 1: Towards populating spreadsheets with LLMs.



SCCS/1436/11

Scientific Committee on Consumer Safety
SCCS

OPINION ON
Basic Red 51
COLIPA n° B116

The SCCS adopted this opinion at its 12th plenary meeting
of 20 September 2011

SCCS/1436/11

Opinion on Basic Red 51

About the Scientific Committees
Three independent non-food Scientific Committees provide the Commission with the scientific advice it needs when preparing policy and proposals relating to consumer safety, public health and the environment. The Committees also draw the Commission's attention to the new or emerging problems which may pose an actual or potential threat.
They are: the Scientific Committee on Consumer Safety (SCCS), the Scientific Committee

Repeated Dose Toxicity

Study 1

Guideline: OECD Guideline 407 Species/strain: Wistar Hanlbm (SPF) rat Group Size: 10 rats per sex: control and the high dose 5 rats per sex: mid and low doses Test material: MIP 2985 Batch: CGF-F016740/0018 Purity: 98% Dose: 15, 50 and 150 mg/kg bw Treatment Period: 28 days Observ. period: 14 days GLP: in compliance MIP 2985 was administered in feed for 28 days at theoretical dose levels of 15, 50 and 150 mg/kg body weight/day while the control group received the normal diet. The corresponding effective daily intake for males was 12.25, 39.59 and 135.46 mg/kg body weight, and for females, 13.16, 40.50 and 132.54 mg/kg body weight for low, mid and high dose groups respectively. A 14-day recovery period was allowed for 5 animals per sex in the control and high dose groups. The animals were examined for clinical signs daily and checked twice daily for mortality. Food

Study 2

Guideline: OECD Guideline 402 Species/strain: Albino Guinea pig, Ibm: GOHI, SPF Group Size: 4 males and 4 females Test material: MIP 2985 Batch: CGF-F016740/0018 Purity: 98% Dose: 1.0, 0.5, 0.1%w/w Observ. period: 14 days GLP: in compliance A 14-day repeated dose dermal toxicity study to assess the cumulative irritation potential with MIP 2985 was applied daily at concentrations of 1.0, 0.5 and 0.1% w/w in double-distilled water. The skin was shaved prior to the study and 2 circular areas of 7 cm² were marked. The animals were shaved four times during the first week and 3 times in the second week. A depilatory cream was used on Day 15 after the final application. One male and female served as controls and received double distilled water. No grading scores were recorded from test day 2-14, as no depilation was used during this period. After depilation on test day 15, no skin reaction was

dt-0?&scraped=true

← sccs_o_067/rdt-0

Guideline: OECD Guideline 407
Species/strain: Wistar Hanlbm (SPF) rat
Group Size: 10 rats per sex: control and the high dose
5 rats per sex: mid and low doses
Test material: MIP 2985
Batch: CGF-F016740/0018

Select Endpoints

Search endpoints

- Reliability
- Endpoints
- Endpoints/Observations
- Test-Method/Test condition
- Test-Method/Test-Substance
- Endpoints/Pathology/Histopathology

Test-Substance

ChatGPT Prompt

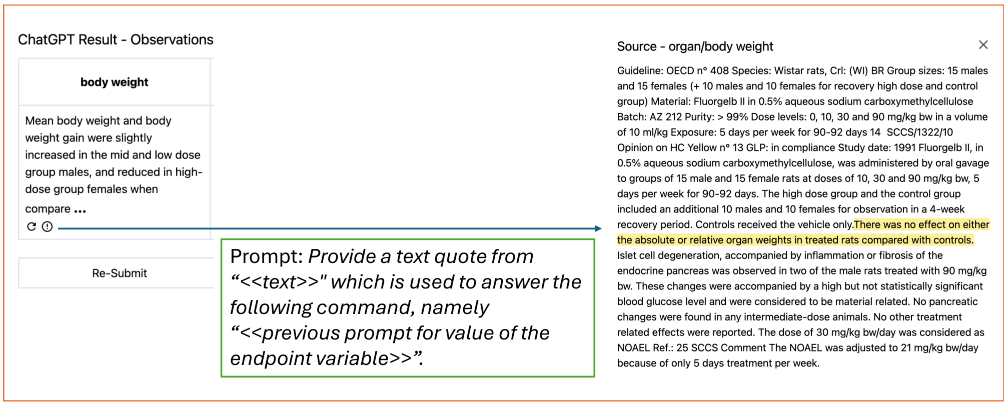
Query ChatGPT

ChatGPT Result - Observations

body weight	food/water consumption	ophtalmological examination	organ/body weight	description Observations
Mean body weight and body weight gain were slightly increased in the mid and low dose group males, and reduced in high-dose group females when compare ... ↻ ⓘ	The relative food consumption was similar in all groups. ↻ ⓘ	- ↻ ⓘ	Organ weights and organ to body weight ratios were higher in the mid-dose group males, but were attributed to the higher terminal body weights of the ... ↻ ⓘ	The animals were examined for clinical signs daily and checked twice daily for mortality. Food consumption and body weight were recorded once pre-test ... ↻ ⓘ

Figure 2: Generating (parts) of a CSV file with LLMs.

column headers are grouped under categories. A user can select one or more such categories. Initial testing has quickly shown that the LLM in our experiment, GPT-4, struggled to generate a coherent CSV with many columns. We generate the following prompt for each column: *"Find the value for the following variable "«column name»" based on the category "«category name»" in the following text "«text»". If you can't find the answer in the text, respond with "-". Don't include any commentary text!"*. The result of which is shown in (3).



ChatGPT Result - Observations

body weight

Mean body weight and body weight gain were slightly increased in the mid and low dose group males, and reduced in high-dose group females when compare ...

Prompt: *Provide a text quote from "<<text>>" which is used to answer the following command, namely "<<previous prompt for value of the endpoint variable>>".*

Source - organ/body weight

Guideline: OECD n° 408 Species: Wistar rats, Crl: (WI) BR Group sizes: 15 males and 15 females (+ 10 males and 10 females for recovery high dose and control group) Material: Fluorgelb II in 0.5% aqueous sodium carboxymethylcellulose Batch: AZ 212 Purity: > 99% Dose levels: 0, 10, 30 and 90 mg/kg bw in a volume of 10 ml/kg Exposure: 5 days per week for 90-92 days 14 SCCS/1322/10 Opinion on HC Yellow n° 13 GLP: in compliance Study date: 1991 Fluorgelb II, in 0.5% aqueous sodium carboxymethylcellulose, was administered by oral gavage to groups of 15 male and 15 female rats at doses of 10, 30 and 90 mg/kg bw, 5 days per week for 90-92 days. The high dose group and the control group included an additional 10 males and 10 females for observation in a 4-week recovery period. Controls received the vehicle only.There was no effect on either the absolute or relative organ weights in treated rats compared with controls. Islet cell degeneration, accompanied by inflammation or fibrosis of the endocrine pancreas was observed in two of the male rats treated with 90 mg/kg bw. These changes were accompanied by a high but not statistically significant blood glucose level and were considered to be material related. No pancreatic changes were found in any intermediate-dose animals. No other treatment related effects were reported. The dose of 30 mg/kg bw/day was considered as NOAEL Ref.: 25 SCCS Comment The NOAEL was adjusted to 21 mg/kg bw/day because of only 5 days treatment per week.

Figure 3: Generating (parts) of a CSV file with LLMs.

Domain experts can recompute the whole CSV table by resubmitting the prompts or the value of one single cell. Domain experts can thus engage with cells multiple times. A promising feature in the prototype is a button prompting the LLM to point to the part of the text that was used to fill in one of the columns. An example is shown in Figure 3. This feature could assist the project in ensuring the data entered in the CSV is authoritative.

4. Exploratory Results

While no user studies have been conducted yet, we deem this approach worthwhile to investigate, given the initial exploratory results. One of the co-authors, a domain expert, found the retrieved information to be often coherent, though experiments with additional domain experiments are warranted. During this study, we noticed that the prompts generated using the column headers sometimes misled the LLM. This was because the column header was ambiguous. This was partly remedied by including the information on the category (e.g., the observed effects of a compound, which are represented under ”Observations” containing five headers, as shown in Figure 2). We plan, however, to investigate specific prompts for each column header, which can be provided to the assistant.

Our current prototype also does not keep track of past interactions; each prompt is executed in a new session. Additional experiments should investigate this impact. What, in our opinion, is more interesting to explore is the use of one-shot or few-shot training. We currently employ zero-shot training with remarkable results. Given the heterogeneity of the Safety Evaluation Opinions, we wonder whether a few-shot approach would yield better results.

5. Conclusions and Future Work

LLMs have been used to generate KGs, but state-of-the-art has shown some challenges with hallucinations and the validity and well-formedness of the KG. We wanted to test whether the generation of CSV would render KG generation more efficient and ensure domain-expert involvement. The advantages are twofold: CSV is an easier and more commonplace data structure, and domain experts are more adept at manipulating spreadsheets. An initial exploration of this approach makes us believe it is worthwhile to investigate.

We developed a prototype that generates CSV based on prompts, which users can copy into a spreadsheet. Subsequently, these spreadsheets are transformed into RDF with RML. It is important to note that the current approach would work for KG projects where domain experts use spreadsheets with existing mappings to a KG.

Future work is twofold: exploring different prompting techniques, as described in the previous section, and integrating the prototype into a workflow for domain experts to allow for domain expert validation.

Acknowledgments

The TOXIN project is financially supported by Vrije Universiteit Brussel under Grant IRP19. Some funding came from Cosmetics Europe and the European Chemical Industry Council.

References

[1] A. Sanctorum, J. Riggio, J. Maushagen, S. Sepehri, E. Arnesdotter, M. Delagrange, J. De Kock, T. Vanhaecke, C. Debruyne, O. De Troyer, End-user engineering of ontology-based knowledge bases, Behaviour & Information Technology 41 (2022) 1811–1829.

[2] C. Debruyne, G. Munnelly, L. Kilgallon, D. O'Sullivan, P. Crooks, Creating a Knowledge Graph for Ireland's Lost History: Knowledge Engineering and Curation in the Beyond 2022 Project, ACM Journal on Computing and Cultural Heritage 15 (2022) 25:1–25:25.

[3] L. Meyer, C. Stadler, J. Frey, N. Radtke, K. Junghanns, R. Meissner, G. Dziwis, K. Bulert, M. Martin, LLM-assisted Knowledge Graph Engineering: Experiments with ChatGPT, in: First Working Conference on Artificial Intelligence Development for a Resilient and Sustainable Tomorrow - AI Tomorrow 2023, Leipzig, Germany, 29-20 June, 2023, Informatik Aktuell, Springer, 2023, pp. 103–115.

[4] A. Randles, D. O'Sullivan, R2[RML]-ChatGPT Framework, in: 5th International Workshop on Knowledge Graph Construction (KGCW 2024) co-located with ESWC 2024, Hersonissos, Greece, May 27, 2024, CEUR Workshop Proceedings, CEUR-WS.org, 2024.

[5] A. Iglesias-Molina, D. Van Assche, J. Arenas-Guerrero, B. De Meester, C. Debruyne, S. Jozashoori, P. Maria, F. Michel, D. Chaves-Fraga, A. Dimou, The RML ontology: A community-driven modular redesign after a decade of experience in mapping heterogeneous data to RDF, in: 22nd International Semantic Web Conference - ISWC 2023, Athens, Greece, November 6-10, 2023, Proceedings, Part II, volume 14266 of *LNCS*, Springer, 2023, pp. 152–175.

[6] M. Hofer, J. Frey, E. Rahm, Towards self-configuring Knowledge Graph Construction Pipelines using LLMs - A Case Study with RML, in: 5th International Workshop on Knowledge Graph Construction (KGCW 2024) co-located with ESWC 2024, Hersonissos, Greece, May 27, 2024, CEUR Workshop Proceedings, CEUR-WS.org, 2024.